

# Smoothing functional data with a Bayesian hierarchical model

Jingjing Yang <sup>\*</sup>, Hongxiao Zhu <sup>†</sup> and Dennis D. Cox <sup>‡</sup>

February 25, 2014

## Abstract

A Bayesian hierarchical model is developed for smoothing functional data. Functional data, with basic data unit being function evaluations (e.g. curves or surfaces) over a continuum, have been frequently encountered in nowadays. While many functional data analysis tools are now available, the issue of simultaneous smoothing is less emphasized. Some methods treat functional data as fully observed while ignoring the measurement noise, others perform smoothing to each functional observation independently thus fail to borrow strength across replications from the same stochastic process. In this paper, we propose a Bayesian hierarchical model to smooth all functional observations simultaneously. The proposed method relies on priors with data-driven hierarchical parameters, which automatically determine the amount of smoothness. It also provides simultaneous estimates for the mean function and covariance. Case studies of simulated and real data demonstrate that this Bayesian method produces more accurate signal estimates and smooth covariance estimate.

*Keywords: functional data, smoothing, Bayesian hierarchical model, Gaussian process, Matern covariance function.*

---

<sup>\*</sup>Department of Statistics, Rice University, Houston, TX 77005, USA, yjingj@gmail.com

<sup>†</sup>Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA, hongxiao@vt.edu

<sup>‡</sup>Department of Statistics, Rice University, Houston, TX 77005, USA, dcox@rice.edu

# 1 Introduction

Since Ramsay and Dalzell (1991) first coined the term “functional data analysis (FDA)”, numerous papers have presented statistical models and tools for functional data. Ramsay and Silverman (2002, Chapter 1) gave a definition of functional data: “a group of data that are defined by functions, most often smooth functions, which could be curves, surfaces, or any function varying over a continuum such as time, spatial location, wavelength, and probability.” For example, the growth records of Swiss boys (Falkner and Falkner, 1960) were discussed in Ramsay (2006) as a typical functional data set, in which the original measurements are the heights of the boys at 29 different ages. Those discretized data can be converted into smooth functions through smoothing techniques.

Many FDA methods are based on methods from multivariate data analysis (MDA). However, FDA methods are connected closely with the domain and expected to have smooth properties in the observations. Müller (2005) stated that “proven methods of MDA form the backbone for some of the prominent techniques of FDA”, such as functional principal components analysis (Silverman, 1996; James et al., 2000; Jiang and Wang, 2010), functional linear regression (Cuevas et al., 2002; Yao et al., 2005; Zhu et al., 2011a), and nonparametric models for functional data (Vieu and Ferraty, 2006; Gardes et al., 2010). Although tremendous FDA methods are motivated from multivariate data analysis (MDA), essential properties — such as smoothness, infinite dimensional consistency, and function evaluation — make FDA a stand-alone field. The most prominent characteristic of functional data is that they are intrinsically infinite dimensional. FDA emphasizes representations of discretized data either through values of the functions on a grid points or coefficients in a basis function expansion, e.g., wavelets. We focus on representing functional data on a grid points in this paper.

Smoothing is an inevitable problem in FDA, in which one aims to estimate smooth functions from finite number of discretized measurements and reduce the effect of noise. Ramsay (2006) introduced many smoothing methods: smoothing splines, locally weighted scatterplot smoothing (lowess), and kernel smoothing with local polynomials. These methods are all based on nonparametric smoothing of individual curve, thus fail to borrow strength across all observations. The more accurate signal estimates we can recover from observed functional data, the better the results that will be produced by further data analyses. For example, Hitchcock et al. (2006) showed that working with smoothed functional data gave more accurate

classification than naively treating the raw data as discretized function evaluations and apply MDA methods.

In this paper, we propose a Bayesian hierarchical model that simultaneously performs smoothing and estimation for the signal mean and covariance. The model assumes Gaussian process likelihood and a Gaussian prior for the underlying smooth curves. We use a Gaussian prior for the mean conditioning on covariance and give a prior on the covariance. A data-driven method is utilized to determine hyper-priors on parameters in the covariance prior and the noise variance. We can use this model to produce estimates of the signal mean, covariance, and smooth signals. Note that our method can be thought of as Bayesian counterpart of Yao et al. (2005), even though our purpose is to provide simultaneous smoothing and covariance estimation rather than eigen-analysis.

There are many challenges in modeling the covariance for the signals. We do not wish to impose assumptions such as stationarity so that we can estimate a general covariance function. As this will be represented by a matrix of its values on a grid of points, we need a prior for the covariance matrix. Many Bayesian models with different prior choices and structure assumptions for the covariance matrix have been proposed. For example, Daniels and Kass (1999) proposed a set of hierarchical priors that produce posterior shrinkage toward a specified structure, e.g., a diagonal matrix; Armstrong et al. (2009) developed an approach by using a prior that was a mixture over all decomposable graphs; Chan and Jeliaskov (2009) built their model upon a decomposition of positive definite matrices to deal with the restriction of positive definite for covariance matrix. These assumed structures and priors of the covariance matrix limit the posterior covariance samples to a smaller space than the space to which they should belong. To avoid such limitations, one will prefer to impose a structure on the hyper-parameter, e.g., on the scale matrix  $\Psi$  in the conjugate prior distribution of the covariance matrix (Inverted -Wishart), as suggested by Chen (1979) who imposed an eigen-decomposed structure on  $\Psi$ .

We seek a prior on the covariance matrix that is consistent with a prior on the space of covariance functions. To this end, we follow a proposal in Zhu et al. (2011b) and use an Inverted Wishart prior with scale matrix  $\Psi$  which is given by a smooth covariance function evaluated at the grid points, and the degrees of freedom depends on the number of grid points. We use a Matern covariance function (Whittle, 1954; Matérn et al., 1960) to give the scale matrix. This has computational advantages as the Inverted Wishart is a conjugate prior for the Gaussian sampling model.

Bayesian estimates are computed via a Markov Chain Monte Carlo

(MCMC) algorithm composed of Gibbs Sampler (Geman and Geman, 1984) and Metropolis-Hastings algorithm (Hastings, 1970). Pilot simulation studies showed that our model was very sensitive to the choices of hyper-priors. We decided to obtain data-driven values for those sensitive hierarchical parameters through an heuristic empirical Bayes method (Casella, 1985).

Our simulation studies showed that the smoothed signal estimates from our Bayesian model were closer to the true signals, than the estimates from splines smoothing, lowess, and kernel smoothing methods. This indicates that by borrowing strength across signal plus noise observations, we improve the individual signal estimates. The Bayesian estimates of the mean and covariance are smoother than the sample mean and sample covariance as well. Since the Bayesian estimate of the covariance is more accurate and smoother, we can generalize the covariance matrix more accurately to a covariance function. Real case studies also suggest that our Bayesian model indeed generates smooth estimates for the signals, signal mean, and covariance.

This article is organized as follows. In Section 2, we give details of the Bayesian hierarchical model and the heuristic empirical Bayes approach to determine hyper-priors. The MCMC algorithm is described in Section 3. Sections 4 and 5 present simulation and real case studies along with the corresponding Bayesian estimates, respectively. Finally, we end with a discussion in Section 6.

## 2 Hierarchical Bayesian model

In this Section, we develop a Bayesian hierarchical model for smoothing functional data, which are assumed as trajectories from a Gaussian process (GP)  $\{Z(t), t \in \mathcal{T}\}$ , with mean function  $\mu(t)$  and covariance function  $\Sigma(s, t) = E[(Z(s) - \mu(s))(Z(t) - \mu(t))]$ . This Gaussian process assumption is generalizable once we have the model developed, that is, the Gaussian likelihood can be generalized for functional data from any other stochastic processes.

### 2.1 Model assumptions

Suppose the functional data are composed of  $n$  independent trajectories from a Gaussian process, where the  $i^{th}$  trajectory has  $p_i$  observed values on the grid  $\mathbf{t}_i = \{t_{i1}, \dots, t_{ip_i}\}$ ,  $i = 1, \dots, n$ . Let  $Y_i$  denotes the  $i^{th}$  observed

trajectory, then the *Signal + Noise* observation model becomes

$$Y_i(t_{ij}) = Z_i(t_{ij}) + \epsilon_{ij}; \quad t_{ij} \in \mathcal{T}, \quad i = 1, \dots, n, \quad j = 1, \dots, p_i, \quad (1)$$

where

- $Z_i(\cdot)$  denotes the true independent and identically distributed (i.i.d.) signals from  $GP(\mu(\cdot), \Sigma(\cdot, \cdot))$ ;
- the noise term  $\epsilon_{ij}$  is assumed to be i.i.d. from a Normal distribution  $N(0, \sigma_\epsilon^2)$  and independent of the signals.

We assume the observed values are on the same grid  $\mathbf{t} = \{t_1, \dots, t_p\}$  for all observations. This assumption of common grid greatly simplifies the computation. In the following context, we denote the  $p \times 1$  vector  $Z_i(\mathbf{t})$  by  $\mathbf{Z}_i$ ,  $Y_i(\mathbf{t})$  by  $\mathbf{Y}_i$ ,  $\mu(\mathbf{t})$  by  $\boldsymbol{\mu}$ ; and denote the  $p \times p$  covariance matrix  $\Sigma(\mathbf{t}, \mathbf{t}')$  by  $\boldsymbol{\Sigma}$ . Then  $\mathbf{Z}_i | (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ,  $i = 1, \dots, n$ , are i.i.d. from multivariate normal distribution  $MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ; and  $\mathbf{Y}_i | (\mathbf{Z}_i, \sigma_\epsilon^2)$ ,  $i = 1, \dots, n$ , are i.i.d. from  $MN(\mathbf{Z}_i, \sigma_\epsilon^2 \mathbf{I})$ , where  $\mathbf{I}$  is the  $p \times p$  identity matrix.

Now we need to consider reasonable prior distributions for the parameters in model (1). Since conjugate priors are preferable in Bayesian analysis, the initial Bayesian model is assumed to have the following partially conjugate priors for  $\sigma_\epsilon^2$ ,  $\boldsymbol{\mu}$ , and  $\boldsymbol{\Sigma}$ :

- $\sigma_\epsilon^2 \sim \text{Inverse-Gamma distribution } IG(a_\epsilon, b_\epsilon)$ , i.e.,  $\gamma_\epsilon = 1/\sigma_\epsilon^2 \sim \text{Gamma distribution } G(a_\epsilon, b_\epsilon)$ ;
- $\boldsymbol{\mu} | \boldsymbol{\Sigma} \sim MN(\boldsymbol{\mu}_0, (1/c)\boldsymbol{\Sigma})$ ;
- $\boldsymbol{\Sigma} \sim \text{Inverted-Wishart distribution } IW(\boldsymbol{\Psi}, \delta)$ , i.e., the precision matrix  $\mathbf{K} = \boldsymbol{\Sigma}^{-1} \sim \text{Wishart distribution } W(\boldsymbol{\Psi}^{-1}, \delta)$ .

$G(a, b)$  distribution has probability density function (PDF)

$$f(\gamma) = b^a \frac{1}{\Gamma(a)} \gamma^{a-1} \exp\{-\gamma b\}.$$

The density function of  $W(\boldsymbol{\Psi}^{-1}, \delta)$  is defined in Dawid (1981) as

$$f(\mathbf{K}) = \frac{|\boldsymbol{\Psi}|^{(\delta+p-1)/2}}{2^{(\delta+p-1)p/2} \Gamma_p((\delta+p-1)/2)} |\mathbf{K}|^{(\delta-2)/2} \exp\left\{-\frac{1}{2} \text{trace}(\boldsymbol{\Psi} \mathbf{K})\right\}, \quad (2)$$

where the scale matrix  $\boldsymbol{\Psi}$  is symmetric and positive definite;  $\delta > 2$  (degrees of freedom is  $\delta + p - 1$ );  $\Gamma_p((\delta + p - 1)/2)$  is the multivariate gamma function defined as  $\pi^{p(p-1)/4} \prod_{i=1}^p \Gamma((\delta + p - i)/2)$ . This Wishart distribution is consistent under finite marginalization. It is justified to use prior  $IW(\boldsymbol{\Psi}, \delta)$  with  $\delta > 4$  for functional covariance, as in Proposition 2.1.

**Proposition 2.1** *Let  $\mathbb{T} \subseteq \mathbb{R}$  be a compact set,  $\delta > 4$ , and  $\Phi : \mathbb{T} \times \mathbb{T} \rightarrow \mathbb{R}$  is a symmetric and positive definite mapping, i.e., any evaluated matrix of  $\Phi$  on a grid  $\{t_1, \dots, t_n\} \times \{t_1, \dots, t_n\} \subseteq \mathbb{T} \times \mathbb{T}$  is symmetric and positive definite. Then there is a probability measure on the space of function mapping  $\mathbb{T} \times \mathbb{T} \rightarrow \mathbb{R}$ , such that if  $V(t, s)$  is drawn from this distribution, then the random  $n \times n$  matrix  $\{V(t_i, t_j) : 1 \leq i, j \leq n\}$  follows a  $IW(\Phi(t_i, t_j) : 1 \leq i, j \leq n, \delta)$  distribution on the grid  $\{t_1, \dots, t_n\} \times \{t_1, \dots, t_n\} \subseteq \mathbb{T} \times \mathbb{T}$ .*

**Proof** The proof of this proposition follows immediately from LEMMA2 in the Appendix of Zhu et al. (2011b).

With  $IW(\Psi, \delta)$  prior, the smaller the  $\delta$  is, the less informative the prior is, but the more likely we would encounter the numerical singularity problem for the covariance posterior-samples. One must take care to choose  $\delta$  to avoid such numerical problems. For example, we took  $\delta = 32$  in the simulation studies and  $\delta = 25$  in the real case studies. The scale matrix  $\Psi$  should be evaluations of a smooth covariance function at the grid points. We impose the Matern covariance structure on  $\Psi$ ,

$$\Psi_{ij} = \text{Matern}(|t_i - t_j|; \text{order} = \nu, \text{length.scale} = \rho, \text{signal.variance} = \sigma^2),$$

where the *Matern* function is defined as

$$\text{Matern}(d; \nu, \rho, \sigma^2) = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left( \sqrt{2\nu} \frac{d}{\rho} \right)^\nu K_\nu \left( \sqrt{2\nu} \frac{d}{\rho} \right), \quad (3)$$

and  $K_\nu(\cdot)$  is the modified Bessel function of the second kind. Then the prior distribution of  $\Sigma$  is  $IW(\Psi = \sigma_s^2 \mathbf{A}, \delta)$ , with  $\mathbf{A}_{ij} = \text{Matern}(|t_i - t_j|; \nu, \rho, \sigma^2 = 1)$ .

Properties of the Matern covariance function are given in Stein (1999). We choose  $\nu > 2$  so that the signals from a Gaussian process with this covariance will be twice differentiable. It is also convenient to take  $\nu$  to be an integer plus  $1/2$ , since then the Matern covariance function will have a closed-form expression. Hence, we take  $\nu = 2.5$ , and then  $\{\mathbf{A}_{ij}; i, j = 1, \dots, p\}$  is given by

$$\mathbf{A}_{ij} = \left( 1 + \frac{\sqrt{5}|t_i - t_j|}{\rho} + \frac{5(|t_i - t_j|)^2}{3\rho^2} \right) \exp \left( -\frac{\sqrt{5}|t_i - t_j|}{\rho} \right). \quad (4)$$

Instead of using parameters  $(\rho, \nu)$  in our Bayesian framework, we transform  $(\rho, \nu)$  to  $\theta = (\theta_1, \theta_2) = (\rho/\sqrt{2\nu}, \nu)$  as in Stein (1999) for

computational simplification, where the value of  $\theta_1$  will be controlling the smoothness of signals and covariance. The Matern covariance function (3) can also be written as

$$\text{Matern}(d; \boldsymbol{\theta}, \sigma^2) = \sigma^2 \frac{1}{\Gamma(\theta_2) 2^{\theta_2-1}} \left( \frac{d}{\theta_1} \right)^{\theta_2} K_{\theta_2} \left( \frac{d}{\theta_1} \right).$$

The following mutually independent hyper-priors are assumed for  $\sigma_s^2$  and  $\boldsymbol{\theta}$ :

- $\sigma_s^2 \sim G(a_s, b_s)$ ;
- $\theta_1 \sim IG(a_{\theta_1}, b_{\theta_1})$ , i.e.,  $\gamma_{\theta_1} = 1/\theta_1 \sim G(a_{\theta_1}, b_{\theta_1})$ ;
- $\theta_2 = 2.5$ .

## 2.2 Heuristic empirical Bayes approach

Now an important issue is how to choose proper values for the hierarchical parameters  $(c, \boldsymbol{\mu}_0, a_\epsilon, b_\epsilon, a_s, b_s, a_{\theta_1}, b_{\theta_1})$ . A pilot simulation study suggested that the Bayesian estimates were quite sensitive to  $(\boldsymbol{\mu}_0, a_\epsilon, b_\epsilon, a_s, b_s, a_{\theta_1}, b_{\theta_1})$ , but not to  $(c, \theta_2)$ . We take  $c = n$ , considering the covariance of the signal average will be scaled by the reciprocal of the number of observations  $n$ . The lowess-smoothed sample mean (with span proportion 0.5) is taken as  $\boldsymbol{\mu}_0$ .

Values for  $(a_\epsilon, b_\epsilon, a_s, b_s, a_{\theta_1}, b_{\theta_1})$  are to be decided by an heuristic empirical Bayes approach, whose details are described in the Appendix. First, we find empirical estimates for  $\{\gamma_\epsilon, \sigma_s^2, \gamma_{\theta_1}\}$ :

- $\hat{\gamma}_\epsilon$  is given by a differencing technique as in equation (6) (Von Neumann, 1941);
- $\hat{\sigma}_s^2$  is a moment estimator obtained by plugging  $\hat{\gamma}_\epsilon$  in the signal plus noise covariance formula as in equation (7);
- $\hat{\gamma}_{\theta_1}$  is solved by matching one value of the Matern correlation with data-based estimate of a stationary autocorrelation.

Then we determine the hyper-priors, respective Gamma distributions of  $\{\gamma_\epsilon, \sigma_s^2, \gamma_{\theta_1}\}$ , by setting the empirical estimates to be the expectations of the corresponding Gamma distributions and choosing a multiple of the estimates as the corresponding variances.

### 2.3 Conditional posteriors

Under the above model, the joint posterior density function is proportional to the joint likelihood function, the product of conditional prior PDFs

$$f(\mathbf{Z}, \boldsymbol{\mu}, \mathbf{K}, \gamma_\epsilon, \sigma_s^2, \boldsymbol{\theta} | \mathbf{Y}) \propto f(\mathbf{Y} | \mathbf{Z}, \gamma_\epsilon) f(\gamma_\epsilon) f(\mathbf{Z} | \boldsymbol{\mu}, \mathbf{K}) f(\boldsymbol{\mu} | \mathbf{K}) f(\mathbf{K} | \sigma_s^2, \boldsymbol{\theta}) f(\sigma_s^2) f(\boldsymbol{\theta}).$$

There are many conditional independencies we can explore for Gibbs Sampler steps in the MCMC algorithm. From the Bayesian framework developed in Section 2.1, one can observe that  $(\mathbf{Z}, \gamma_\epsilon)$  depends on everything else through  $(\mathbf{Y}, \boldsymbol{\mu}, \mathbf{K})$  only. Given  $\mathbf{Z}$ , the conditional posterior distribution for  $(\boldsymbol{\mu}, \mathbf{K}, \sigma_s^2, \boldsymbol{\theta})$  can be derived free of  $(\mathbf{Y}, \gamma_\epsilon)$ . This strategy leads to conjugate conditional posteriors for all parameters except  $\boldsymbol{\theta}$ . We only present an outline of the results and conditional density functions, because all derivations are trivial.

Given the observed data  $\mathbf{Y}$  and parameters  $(\boldsymbol{\mu}, \mathbf{K}, a_\epsilon, b_\epsilon)$ , the joint conditional posterior distribution of  $(\mathbf{Z}, \gamma_\epsilon)$  becomes

$$f(\mathbf{Z}, \gamma_\epsilon | \mathbf{Y}, \boldsymbol{\mu}, \mathbf{K}, a_\epsilon, b_\epsilon) \propto f(\mathbf{Y} | \mathbf{Z}, \gamma_\epsilon) f(\gamma_\epsilon | a_\epsilon, b_\epsilon) f(\mathbf{Z} | \boldsymbol{\mu}, \mathbf{K}).$$

Then the full conditional posterior distributions for  $\mathbf{Z}_i, i = 1, \dots, n$  are i.i.d. with multivariate normal distribution

$$\mathbf{Z}_i | \mathbf{Y}_i, \boldsymbol{\mu}, \mathbf{K}, \gamma_\epsilon \sim MN \left( (\gamma_\epsilon \mathbf{I} + \mathbf{K})^{-1} (\gamma_\epsilon \mathbf{Y}_i + \mathbf{K} \boldsymbol{\mu}), (\gamma_\epsilon \mathbf{I} + \mathbf{K})^{-1} \right),$$

while the full conditional posterior distribution for  $\gamma_\epsilon$  is given by

$$\gamma_\epsilon | \mathbf{Y}, \mathbf{Z}, a_\epsilon, b_\epsilon \sim G \left( a_\epsilon + \frac{np}{2}, b_\epsilon + \frac{1}{2} \text{trace}[(\mathbf{Y} - \mathbf{Z})(\mathbf{Y} - \mathbf{Z})^T] \right).$$

The joint conditional posterior distribution of  $(\boldsymbol{\mu}, \mathbf{K}, \sigma_s^2, \boldsymbol{\theta})$  depends on  $\mathbf{Z}$  and hierarchical parameters  $(a_s, b_s, a_{\theta_1}, b_{\theta_1})$ , given by

$$f(\boldsymbol{\mu}, \mathbf{K}, \sigma_s^2, \boldsymbol{\theta} | \mathbf{Z}, a_s, b_s, a_{\theta_1}, b_{\theta_1}) \propto f(\mathbf{Z} | \boldsymbol{\mu}, \mathbf{K}) f(\boldsymbol{\mu} | \mathbf{K}) f(\mathbf{K} | \sigma_s^2, \boldsymbol{\theta}) f(\sigma_s^2 | a_s, b_s) f(\boldsymbol{\theta} | a_{\theta_1}, b_{\theta_1}).$$

Conditioning on  $\mathbf{Z}$ , we can treat it as  $\mathbf{Z}$  were the data. Then the first factor on the right side above is like the likelihood and the subsequent factors follow from the priors for  $\boldsymbol{\mu}$  and  $\mathbf{K}$ . Similarly, by Bayes rule, given  $\mathbf{Z}$  and  $\Psi(\sigma_s^2, \boldsymbol{\theta})$ , the respective conditional posterior distributions of  $\boldsymbol{\mu}, \mathbf{K}$  are

$$\boldsymbol{\mu} | \mathbf{K}, \mathbf{Z} \sim MN \left( \frac{1}{n+c} \left( \sum_{i=1}^n \mathbf{Z}_i + c \boldsymbol{\mu}_0 \right), \frac{1}{n+c} \mathbf{K}^{-1} \right),$$



and

$$\mathbf{K}|\boldsymbol{\mu}, \mathbf{Z}, \Psi(\sigma_s^2, \boldsymbol{\theta}) \sim W \left( [(\mathbf{Z} - \boldsymbol{\mu}\mathbf{J})(\mathbf{Z} - \boldsymbol{\mu}\mathbf{J})^T + c\boldsymbol{\mu}\boldsymbol{\mu}^T + \Psi(\sigma_s^2, \boldsymbol{\theta})]^{-1}, n + \delta + p \right),$$

i.e.,

$$\boldsymbol{\Sigma}|\boldsymbol{\mu}, \mathbf{Z}, \Psi(\sigma_s^2, \boldsymbol{\theta}) \sim IW \left( [(\mathbf{Z} - \boldsymbol{\mu}\mathbf{J})(\mathbf{Z} - \boldsymbol{\mu}\mathbf{J})^T + c\boldsymbol{\mu}\boldsymbol{\mu}^T + \Psi(\sigma_s^2, \boldsymbol{\theta})], n + \delta + p \right),$$

where  $\mathbf{J} = (1, \dots, 1)_{1 \times p}$ ,  $\Psi(\sigma_s^2, \boldsymbol{\theta}) = \sigma_s^2 \mathbf{A}(\boldsymbol{\theta})$ , and  $\mathbf{A}_{ij}(\boldsymbol{\theta})$  is given by equation (4) with  $\nu = \theta_2$ ,  $\rho = \theta_1 \sqrt{2\theta_2}$ .

Given  $\mathbf{K}$ , the respective conditional posterior distributions of  $\sigma_s^2, \gamma_{\theta_1}$  are derived as

$$\sigma_s^2|\mathbf{K}, \boldsymbol{\theta}, a_s, b_s \sim G \left( a_s + \frac{mp}{2}, \quad b_s + \frac{1}{2} \text{trace}[\mathbf{A}(\boldsymbol{\theta})\mathbf{K}] \right)$$

and

$$f(\gamma_{\theta_1}|\mathbf{K}, \sigma_s^2, a_{\theta_1}, b_{\theta_1}, \theta_2) \propto |\mathbf{A}(\boldsymbol{\theta})|^{m/2} \exp \left\{ -\frac{1}{2} \text{trace}[\mathbf{A}(\boldsymbol{\theta})\mathbf{K}] \right\} \gamma_{\theta_1}^{a_{\theta_1}-1} \exp(-\gamma_{\theta_1} b_{\theta_1}).$$

Since  $f(\gamma_{\theta_1}|\mathbf{K}, \sigma_s^2, a_{\theta_1}, b_{\theta_1}, \theta_2)$  is not a known distribution, we use a Metropolis-Hastings step to sample posterior-samples of  $\gamma_{\theta_1}$ .

### 3 Markov Chain Monte Carlo algorithm

With the Bayesian framework, we can obtain posterior samples for all parameters through a MCMC algorithm. Here, the MCMC algorithm is a mixture of Gibbs Sampler and Metropolis-Hastings algorithm. We show the MCMC algorithm as follows by taking the  $(k+1)^{th}$  iteration as an example, where the supper script denotes the iteration step in which the parameter value is obtained.

- First, generate the smoothed signal estimates for  $i = 1, \dots, n$  by

$$\mathbf{Z}_i^{(k+1)} \sim MN \left( (\gamma_\epsilon^{(k)} \mathbf{I} + \mathbf{K}^{(k)})^{-1} (\gamma_\epsilon^{(k)} \mathbf{Y}_i + \mathbf{K}^{(k)} \boldsymbol{\mu}^{(k)}), \quad (\gamma_\epsilon^{(k)} \mathbf{I} + \mathbf{K}^{(k)})^{-1} \right).$$

As a result,  $\mathbf{Z}^{(k+1)} = (\mathbf{Z}_1^{(k+1)}, \dots, \mathbf{Z}_n^{(k+1)})$ .

- Second, update the parameters:

$$\gamma_\epsilon^{(k+1)} \sim G \left( a_\epsilon + \frac{np}{2}, \quad b_\epsilon + \frac{1}{2} \text{trace}[(\mathbf{Y} - \mathbf{Z}^{(k+1)})(\mathbf{Y} - \mathbf{Z}^{(k+1)})^T] \right);$$

$$\begin{aligned}
\boldsymbol{\mu}^{(k+1)} &\sim MN\left(\frac{1}{n+c}\left(\sum_{i=1}^n \mathbf{Z}_i^{(k+1)} + c\boldsymbol{\mu}_0\right), \quad \frac{1}{n+c}(\mathbf{K}^{(k)})^{-1}\right); \\
\boldsymbol{\Sigma}^{(k+1)} &\sim IW\left(\mathbf{Q}^{(k)}, n+\delta+p\right), \quad \mathbf{K}^{(k+1)} = (\boldsymbol{\Sigma}^{(k+1)})^{-1}, \text{ where} \\
\mathbf{Q}^{(k)} &= (\mathbf{Z}^{(k+1)} - \boldsymbol{\mu}^{(k+1)}\mathbf{J})(\mathbf{Z}^{(k+1)} - \boldsymbol{\mu}^{(k+1)}\mathbf{J})^T + \\
&\quad c(\boldsymbol{\mu}^{(k+1)} - \boldsymbol{\mu}_0)(\boldsymbol{\mu}^{(k+1)} - \boldsymbol{\mu}_0)^T + (\sigma_s^2)^{(k)}\mathbf{A}(\boldsymbol{\theta}^{(k)}).
\end{aligned}$$

- Finally, update the hierarchical parameters with

$$(\sigma_s^2)^{(k+1)} \sim G\left(a_s + \frac{mp}{2}, \quad b_s + \frac{1}{2}\text{trace}[\mathbf{A}(\boldsymbol{\theta}^{(k)})\mathbf{K}^{(k+1)}]\right);$$

and  $(\theta_1)^{(k+1)}$  sampled by the Metropolis-Hestings algorithm with Log-Normal proposal distribution  $LN(\log(\theta_1^{(k)}), 1)$ , where the proposed value for  $\log((\theta_1)^{(k+1)})$  is distributed normally with mean  $\log(\theta_1^{(k)})$  and standard deviation 1.

For the case studies in Sections 4 and 5, we found that at least 10,000 MCMC iterations with extra 2,000 iterations of burn-in should be implemented, so that the posterior samples could achieve convergence. In each case study, we selected hyper priors by the empirical Bayes approach described in Section 2.2 and Appendix, then ran the MCMC algorithm with starting values  $\{\gamma_\epsilon^{(0)} = \widehat{\gamma}_\epsilon, (\sigma_s^2)^{(0)} = \widehat{\sigma}_s^2, \gamma_{\theta_1}^{(0)} = \widehat{\gamma}_{\theta_1}, \boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}_0, \boldsymbol{\Sigma}^{(0)} = \mathbf{I}_{p \times p}\}$ .

## 4 Simulation studies

### 4.1 Stationary Gaussian process data

For each simulation process, we generated  $n = 100$  stationary functional data curves (true signals) from

$$GP(\mu(t) = 3\sin(4t), \Sigma(s, t) = \text{Matern}(|s - t|; \nu = 5/2, \rho = 1/2, \sigma^2 = 5)), \quad (5)$$

denoted by  $\mathbf{Z}^*$ . All curves were obtained on the same equally spaced grid  $\{t_{i1} = 0, \dots, t_{ip} = 1\}$  with  $p = 31$ . The noise terms  $\{\epsilon_{ij}; i = 1, \dots, 100, j = 1, \dots, 31\}$  were generated i.i.d. from  $N(0, \sigma_\epsilon = \sqrt{5}/2)$ . The observed noisy functional data curves were given by  $\mathbf{Y} = \mathbf{Z}^* + \boldsymbol{\epsilon}$  (first 10 noisy curves are plotted in the top frame of Figure 1).

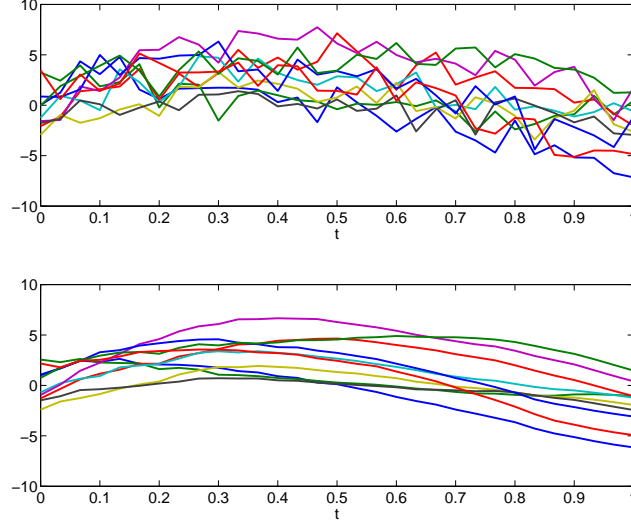
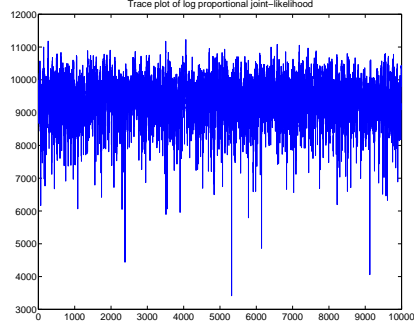


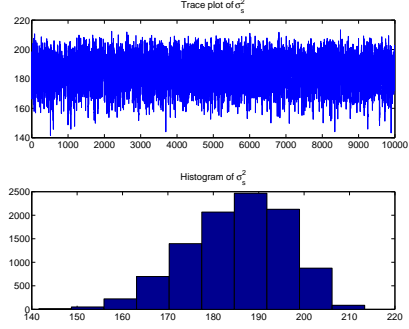
Figure 1: First 10 raw stationary functional curves (top frame), along with the corresponding Bayesian smoothed curves (bottom frame).

First, we calculated  $\widehat{\gamma}_\epsilon = 0.79$ ,  $\widehat{\sigma_s^2} = 156.39$ , and  $\widehat{\gamma_{\theta_1}} = 5.88$ . Next, with  $\{b_\epsilon = 1/3, b_s = 30, b_{\theta_1} = 1/3\}$ , the hierarchical parameter values  $\{a_\epsilon = 0.26, a_s = 4691.8, a_{\theta_1} = 1.94\}$  were determined by the empirical Bayes approach. Finally, we took the corresponding posterior-sample-averages from the final 10, 000 MCMC iterations as the Bayesian estimates for all parameters. Examining the respective posterior-sample trace plots of the log proportional joint-likelihood,  $\sigma_s^2$ ,  $\sigma_\epsilon$ , and  $\gamma_{\theta_1}$  in Figure 2, we can see that the final posterior-samples have all mixed well and achieved convergence.

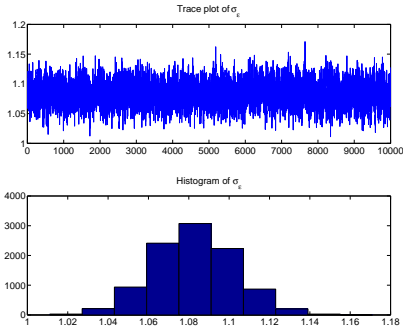
We display the Bayesian signal estimates for the first 10 noisy signals in the bottom frame of Figure 1, 4 groups of selected signals (true/noisy/smoothed) in Figure 3, mean curves in Figure 4a, the plot of the Bayesian estimated variances (diagonal of the estimated covariance) in Figure 4b, and the heat map of the Bayesian estimated correlation in Figure 5b contrasting with the sample estimate in Figure 5a and the true correlation in Figure 5c. It is obvious that the Bayesian signal estimates are smoother and closer to the true signals, with 95% credible intervals covering the true signals. Although the Bayesian estimated signal mean almost overlaps with the sample mean, the 95% credible interval covers the true mean curve. In



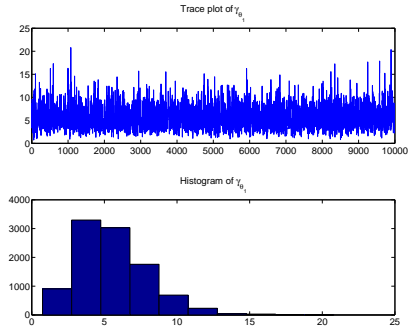
(a)



(b)



(c)



(d)

Figure 2: (a): Trace plot of log proportional joint-likelihood; (b), (c) & (d): trace plots and histograms of the final MCMC posterior-samples for  $\sigma_s^2$ ,  $\sigma_\epsilon$  and  $\gamma_{\theta_1}$ , respectively.

addition, even though the Bayesian estimate of the covariance tracks mainly the sample estimate, it is smoother than the sample estimate and closer to the true covariance.

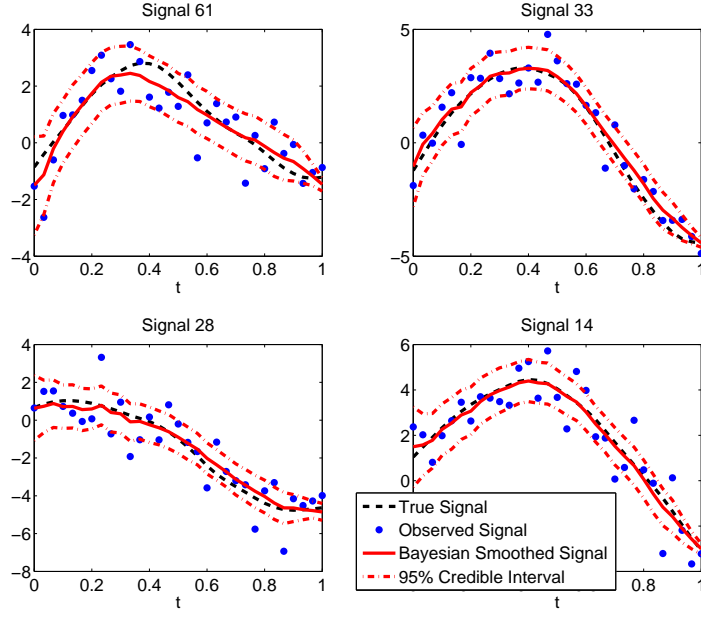


Figure 3: Selected four true/noisy/smoothed curves (along with 95% credible interval), with MSE of the smoothed curve decreasing from the maximum (top left) to the minimum (bottom right).

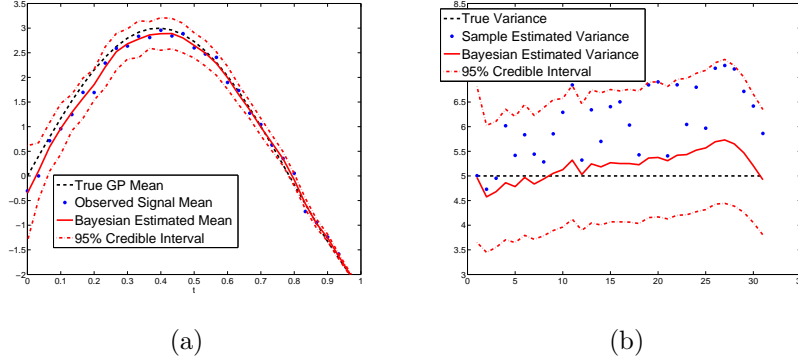


Figure 4: (a): True GP mean, sample mean, and Bayesian estimated signal mean (along with 95% credible interval); (b): plots of the true, sample estimated, and Bayesian estimated variances (along with 95% credible interval).

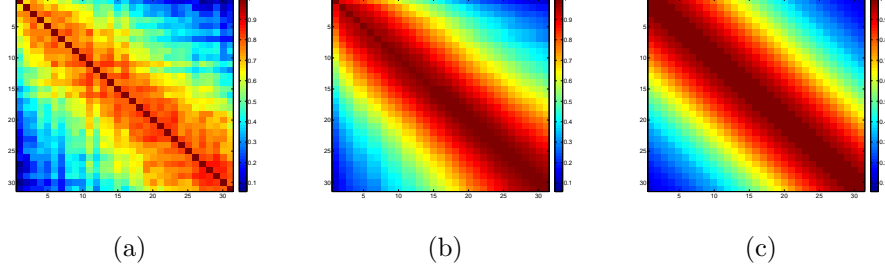


Figure 5: (a): Sample estimated correlation; (b): Bayesian estimated correlation; (c): true correlation (stationary Gaussian data).

We compare our Bayesian signal estimates with the best possible least square estimates and the ones generated by other approaches, such as fitting a cubic smoothing spline, lowess, and kernel smoothing with local polynomials. Recall the observation model (1), the joint distribution of  $\begin{pmatrix} \mathbf{Z}_i \\ \mathbf{Y}_i \end{pmatrix}$  is i.i.d. with  $MN \left[ \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I} \end{pmatrix} \right], i = 1, \dots, n$ . Given  $\mathbf{Y}$ , true mean  $\boldsymbol{\mu}$ , and true covariance  $\boldsymbol{\Sigma}$ , we can obtain the conditional mean and covariance of  $\mathbf{Z}$ :

$$\begin{aligned} E[\mathbf{Z}_i | \mathbf{Y}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}] &= \boldsymbol{\mu} + \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I})^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}); \\ Cov(\mathbf{Z}_i | \mathbf{Y}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I})^{-1}\boldsymbol{\Sigma}, \quad i = 1, \dots, n. \end{aligned}$$

The conditional mean  $E[\mathbf{Z}_i | \mathbf{Y}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}]$  is the best possible least square estimate for the  $i^{th}$  signal, with the variance at each dimension given by the corresponding diagonal value of the conditional covariance  $Cov(\mathbf{Z}_i | \mathbf{Y}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ . We obtained the other smoothing estimates by using R (R Core Team, 2013) functions: “smooth.spline” with smoothing parameter value decided by the default generalized cross-validation (GCV); “lowess” with default smoother span 2/3; and “locpoly( $\dots$ , degree = 1)” from package “KernSmooth” (Wand and Jones, 1995) with bandwidth selected by a plug-in method (R function “dpill”).

we ran the above simulation process for 200 times and took the average mean square error (MSE),  $(1/200) \sum_{k=1}^{200} [(\sum_{i=1}^n \sum_{j=1}^p (\hat{Z}_{ij} - Z_{ij})^2) / (np)]$ , as a criterion. As seen from Tables 1, 2, and 3 about the average MSEs (along with the MSE standard deviations) for signal estimates, correlation estimates, and variance estimates (diagonal of the covariance) obtained from the 200 times of simulation, our Bayesian method produced the most accurate signal estimates and covariance estimate for stationary Gaussian data.

## 4.2 Non-stationary Gaussian data

From the simulation study of stationary Gaussian data, we have seen that the Bayesian method produced good estimates for signals and covariance. Now we explore how this Bayesian method will perform with non-stationary Gaussian data. We generated non-stationary Gaussian data by imposing a nonlinear transformation on the variable  $t$  and a scalar function on the Gaussian process (5) in Section 4.1.

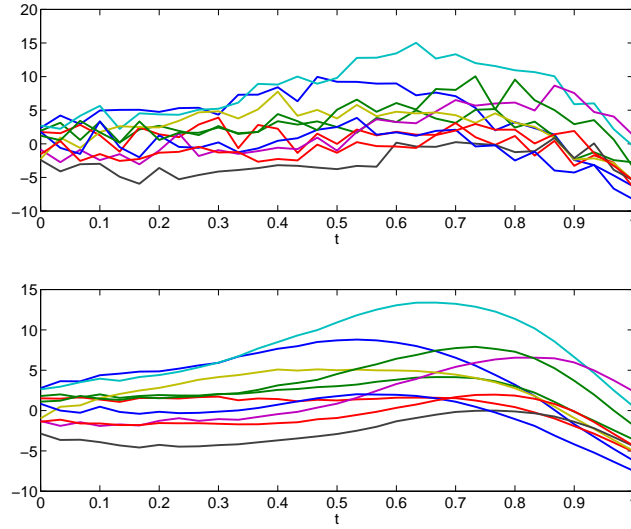


Figure 6: First 10 raw non-stationary functional curves (top frame), along with the corresponding Bayesian smoothed estimates (bottom frame).

Let the nonlinear transformation function be  $\xi(t) = t^2$ ; the scalar function be  $h(t) = 1 + t$ ;  $X_i(t)$  denote a functional trajectory generated from Gaussian process (5). Then the non-stationary true trajectory is given by  $Z_i^*(t) = h(t)X_i(\xi(t))$ , i.e., a trajectory from

$$GP(h(t)3\sin(\xi(t)), \Sigma(s, t) = h(s)h(t)Matern(|\xi(s) - \xi(t)|; \nu = 2.5, \rho = 1/2, \sigma^2 = 5)).$$

For each simulation process, 100 non-stationary functional trajectories were generated on the same equally spaced grid  $\{t_{i1} = 0, \dots, t_{i31} = 1\}$ . Then we obtained the noisy functional data  $\mathbf{Y}$  (the first 10 noisy curves were plotted in the top frame of Figure 6) by adding on noise terms i.i.d. from  $N(0, \sigma_\epsilon = \sqrt{5}/2)$ .

Applying the Bayesian smoothing method, we obtained smoothed signal estimates (estimates for the first 10 noisy signals are shown in the bottom frame of Figure 6, selected 4 groups of true/noisy/smoothed signals are given in Figure 7), the Bayesian estimated mean-signal almost overlapping with the sample mean (Figure 8a), and the estimate of the covariance (diagonal variances are plotted in Figure 8b; the heat map of the correlation is given in Figure 9b). The average MSEs for signal estimates, correlation estimates, and variance estimates (diagonal of the covariance) from 200 times of simulation are presented in Tables 1, 2, and 3, respectively. From those plots and tables, we can see that the Bayesian estimates still possess smoothing property and smallest MSEs for signal and correlation estimates.

Table 1: Average MSEs (along with MSE standard deviations) for signal estimates from cubic-smoothing-splines/lowess/kernel-smoothing/Bayesian in comparison with the best possible least square estimates, which gives a lower bound of the MSE.

Average MSEs (Signal)	Splines	Lowess	Kernel	Bayesian	Lower Bound
Stationary	0.325 ( $\pm 0.0021$ )	0.296 ( $\pm 0.0013$ )	0.275 ( $\pm 0.0262$ )	0.170 ( $\pm 0.0008$ )	0.155 ( $\pm 0.0007$ )
Non-stationary	0.389 ( $\pm 0.0023$ )	1.445 ( $\pm 0.0089$ )	0.371 ( $\pm 0.0016$ )	0.227 ( $\pm 0.0011$ )	0.186 ( $\pm 0.0008$ )

Table 2: Average MSEs (along with MSE standard deviations) for sample estimate and Bayesian estimate of the correlation.

Average MSEs (Correlation)	Sample estimate	Bayesian estimate
Stationary	0.026 ( $\pm 0.0008$ )	0.004 ( $\pm 0.0004$ )
Non-stationary	0.012 ( $\pm 0.0004$ )	0.008 ( $\pm 0.0004$ )

However, from the correlation heat maps in Figure 9, we can see that the assumed  $IW$  prior with stationary scale matrix  $\Psi$  shrank the upper left corner of the Bayesian estimate toward the diagonal. Figure 8b and Table 3 also show that the Bayesian method tends to overestimate the



Table 3: Average MSEs (along with MSE standard deviations) for sample estimate and Bayesian estimate of the variances (diagonal of the covariance).

Average RMSEs (Variances)	Sample estimate	Bayesian estimate
Stationary	2.392 ( $\pm$ 0.0984)	0.576 ( $\pm$ 0.0395)
Non-stationary	4.922 ( $\pm$ 0.2934)	5.535 ( $\pm$ 0.1841)

covariance diagonal values, which is caused by the  $\sigma_s^2$  value in  $\Psi$ . Hence, we could improve the Bayesian covariance estimate by assuming an adaptable non-stationary  $\Psi$  in the  $IW$  prior.

## 5 Applications on real data

### 5.1 Spectroscopy data

We applied the Bayesian smoothing method to the spectroscopy data ( $n = 742$ ), which were measured on patient tissue samples by two multispectral digital colposcope (MDC) devices (Buys et al., 2012) with white light. Those spectroscopy measurements were made to help predict cervical cancer in the early stages. Each measurement was composed of intensity values at emission wavelengths from 375 to 750, which could be treated as functional data on wavelength. Observed raw data were scaled by 3000 such that the intensity values were suitable for computation. Since the main information is contained within the emission wavelength interval (434, 704) and each value is very close to its neighbors, we took values at 91 equally spaced emission wavelengths from (434, 704) with increment 3 as our observations  $\mathbf{Y}$  (10 example raw-signals are plotted in the top frame of Figure 10).

After applying the Bayesian smoothing method to the raw signals, we have smoothed signal estimates (10 example signal-estimates are plotted in the bottom frame of Figure 10), the Bayesian estimate of the signal mean as in Figure 11a, and the Bayesian estimate of the covariance (diagonal variances are plotted in Figure 11b; the heat map of the correlation is shown in Figure 12b). As we can observe from Figure 10, there are some wiggles in the raw curves. Almost all of these small wiggles would be removed by other individual smoothing methods. However, our Bayesian estimated signals kept those wiggles that are repeated among most measurements and possibly caused by systematic errors, so that smoothed signals would not

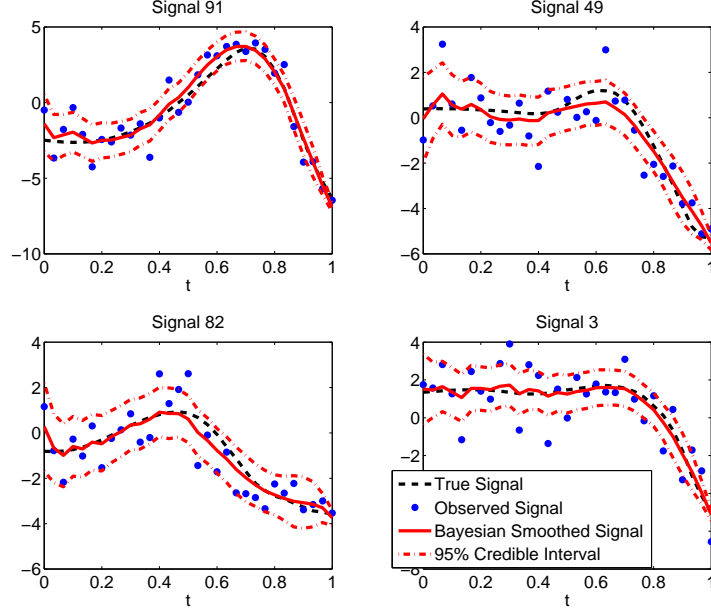


Figure 7: Selected four groups of true/noisy/smoothed curves (along with 95% credible interval), with MSE of the smoothed curve decreasing from the maximum (top left) to the minimum (bottom right)

lose that information.

We use the rest 2/3 data points as a validation data set to compare our Bayesian method with smoothing splines (smoothing parameter decided by GCV), lowess (smoother span 1/10), and kernel smooth (bandwidth selected by R function “dpill”). With fitted smoothing splines for each curve over wavelengths 434 : 3 : 704, we can get prediction values on the rest wavelengths. With the Bayesian, lowess, and kernel estimates over wavelengths 434 : 3 : 704 (1/3 data points), we linearly interpolated each curve estimate to obtain prediction values on the validation wavelengths. Because this spectroscopy data have few noise, we treat the observed values on the validation wavelengths as true data and calculate the MSE per curve for each smoothing method. Table 4 shows the average MSEs of all curves, along with standard deviation in the parentheses.

One can also observe in Figures 11b and 12 that the Bayesian estimated variances do not drop as dramatically as the sample estimated ones in both

Table 4: Average MSEs (along with MSE standard deviations) on the validation data of spectroscopy curves.

	Splines	Lowess	Kernel	Bayesian
Average MSEs	0.0014 ( $\pm 5.9 \times 10^{-5}$ )	0.0064 ( $\pm 4.9 \times 10^{-4}$ )	0.0049 ( $\pm 3.5 \times 10^{-4}$ )	0.0013 ( $\pm 5.5 \times 10^{-5}$ )

ends and the Bayesian estimated correlation are shrunk toward the diagonal, which can be explained by the stationary prior. This real data study gives an example about how our Bayesian method behaves for data that contain a relatively small amount of noise. The Bayesian signal estimates would preserve more information from the original observations, along with a fairly good estimation of the covariance.

## 5.2 Northeast coast wind speed data

In this real case study, we analyzed a group of wind speed data gathered at 26 buoys on the northeast coast of North America. The data are available at the National Buoy Data Center (NBDC) historical data page <http://www.ndbc.noaa.gov/>. We focused on the hourly average wind speed (first 10 curves are shown in the top frame of Figure 13) on January 18, 2005.

As a result of applying our Bayesian smoothing method to the raw data, we have Bayesian signal estimates (estimates of the first 10 curves are shown in the bottom frame of Figure 13; selected 4 groups of observed/smoothed curves are plotted in Figure 14), the Bayesian estimate of the mean curve as in Figure 15a, and the Bayesian estimate of the covariance (diagonal variances are shown in Figure 15b; the heat map of the correlation is given in Figure 16b). From those plots, one can observe clearly that our model has produced smoothed signal estimates and a covariance estimate for real data containing a relatively large amount of noise.

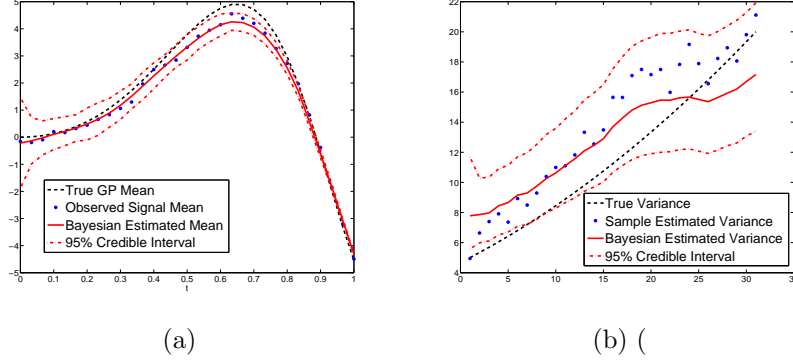


Figure 8: (a): True GP mean, sample mean, and Bayesian estimated signal mean (along with 95% credible interval); (b): plots of the true, sample estimated and Bayesian estimated variances (along with 95% credible interval).

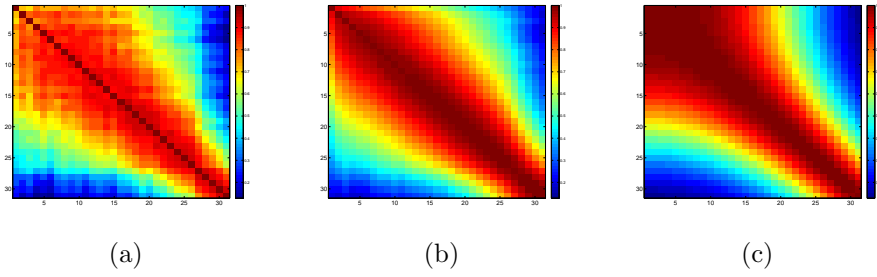


Figure 9: (a): Sample estimated correlation; (b): Bayesian estimated correlation; (c) true correlation (non-stationary Gaussian data).

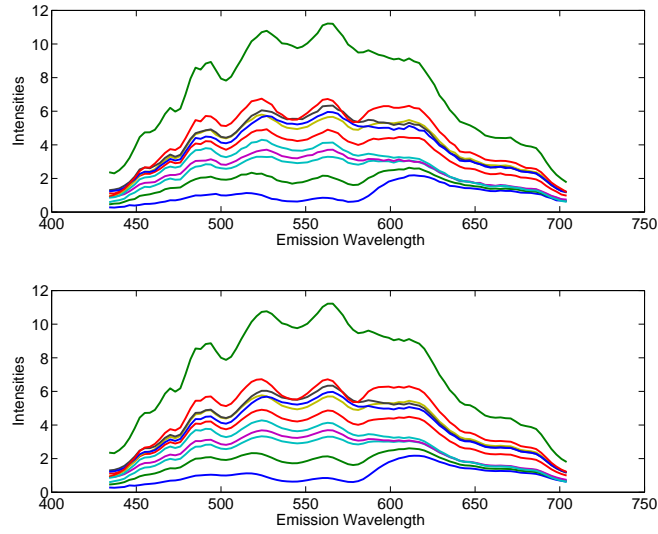


Figure 10: Ten example raw-spectroscopy-curves (top frame), along with the corresponding Bayesian smoothed curves (bottom frame).

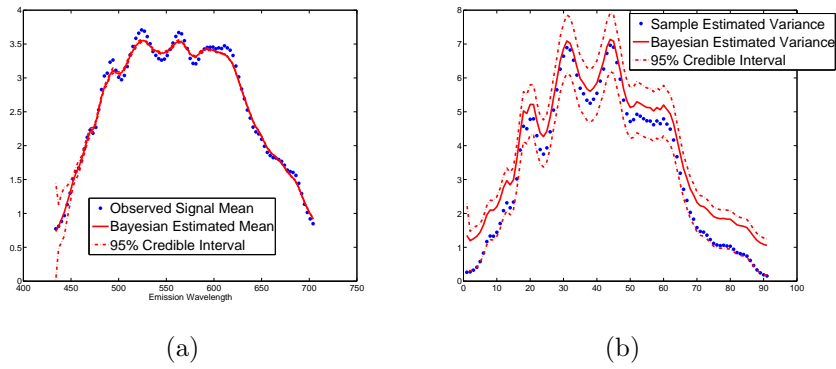


Figure 11: (a): Sample mean and Bayesian estimated signal mean (along with 95% credible interval); (b): plot of the sample estimated and Bayesian estimated variances (along with 95% credible interval).

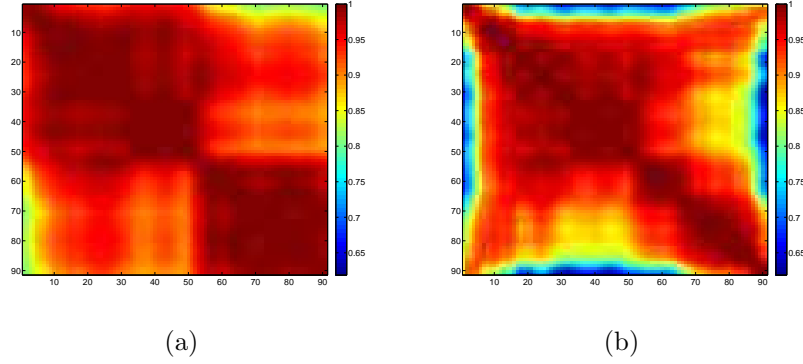


Figure 12: (a): Sample estimated correlation; (b): Bayesian estimated correlation (spectroscopy data).

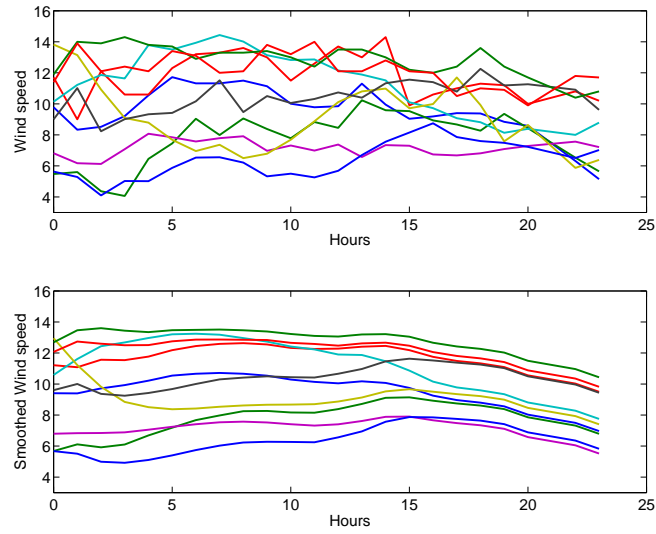


Figure 13: First 10 raw wind speed curves (top frame), along with the corresponding Bayesian smoothed curves (bottom frame).

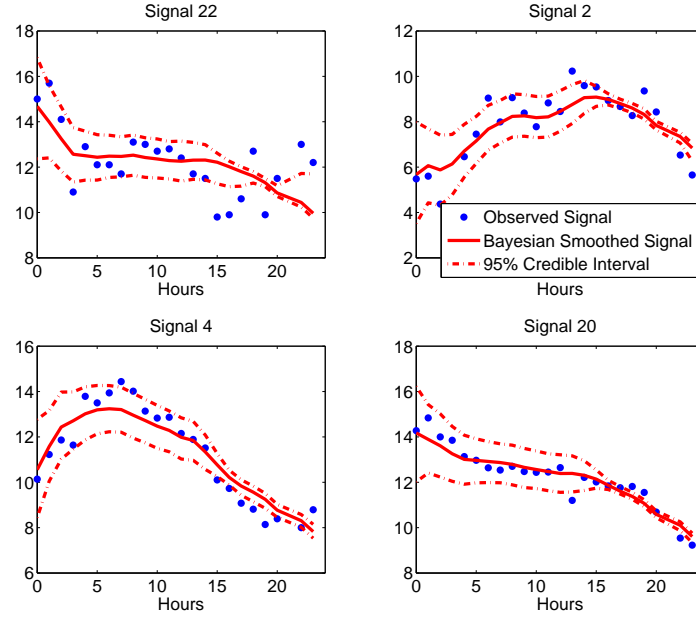
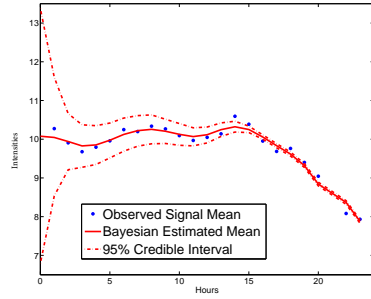
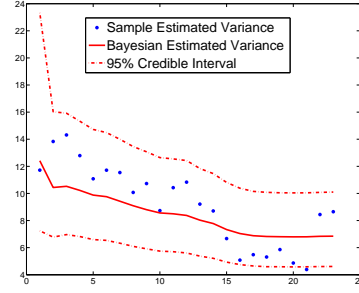


Figure 14: Selected 4 groups of observed/smoothed wind speed curves (along with 95% credible interval), with the mean square distance between the Bayesian estimate and the observation decreasing from the maximum (top left) to the minimum (bottom right)

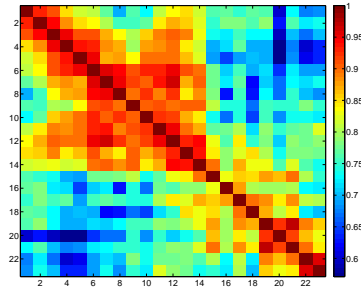


(a)

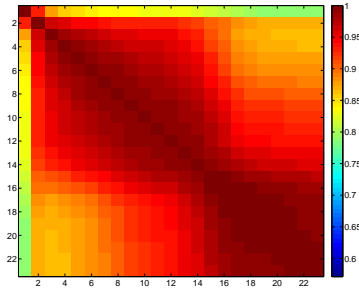


(b)

Figure 15: (a): Sample mean and Bayesian estimated signal mean (along with 95% credible interval); (b): plot of the sample estimated and the Bayesian estimated variances (along with 95% credible interval).



(a)



(b)

Figure 16: (a): Sample estimated correlation; (b): Bayesian estimated correlation (wind speed data).



## 6 Discussion

In this paper, we propose a new Bayesian hierarchical model with an empirical Bayes approach to determine hierarchical parameter values, which can borrow strength across all observations to smooth raw functional data simultaneously and give smooth estimates for the signal mean and covariance. It is common in FDA to assume the observed functions are i.i.d.. Although our Bayesian signal estimates will not be i.i.d. — because each one of them will be generated by using information from the others — they will still be exchangeable. Moreover, the Bayesian smoothed signals are likely to keep features repeated across most observations that may be due to systematic errors, so that less information will be lost in the smoothing process. Since the Bayesian estimates of the signals, signal mean and covariance are discretized values, we can easily interpolate them to smooth functions as needed. One has to be careful when interpolating the covariance matrix, because the covariance function has to be symmetric and positive definite.

Most existing smoothing methods work with individual observation and ignore the fact that the raw data may come from the same stochastic process. Hence, our model will perform better when the observed functional data are trajectories from the same stochastic process, such as in the case studies in Sections 4 and 5. In addition, the Bayesian estimate of the correlation is smoother and more accurate than the sample estimate for both stationary and non-stationary Gaussian data.

Although the Bayesian method tends to shrink the correlation estimate toward the diagonal and over estimate the diagonal of the covariance for non-stationary Gaussian data, we can surely improve the Bayesian estimates by adapting non-stationarity in the inverted Wishart prior of the functional data covariance. There is some literature investigating the problem of how to obtain a nonparametric estimate for non-stationary covariance structure. Sampson and Guttorp (1992) introduced a semiparametric approach to obtain non-stationary spatial covariance structure, and Hall et al. (1994) described a nonparametric kernel method to estimate the covariance function of a stationary stochastic process. We plan to extend our Bayesian method to give a better nonparametric estimate for the non-stationary covariance structure.

One of the limitations in our model is that the number of replications ( $n$ ) has to be larger than or equal to the number of grid points ( $p$ ). Otherwise, we will encounter the issue of non-positive definite scale matrix in the conditional posterior distribution (inverted-Wishart) of the covariance

matrix, as a result of which the posterior covariance samples can not be generated. The future work will also include extending the model to deal with cases with functional data from multiple stochastic processes, different grids for each replication, and  $n < p$ . Even though we are focusing on building an applicable Bayesian model in this paper, we also would like to establish the theoretical properties such as posterior consistency and rates of convergence.

## Appendix: Details of the empirical Bayes approach

With this empirical Bayes approach, we can obtain values for the hierarchical parameters in the corresponding Gamma distributions, starting with empirical estimates for  $\{\gamma_\epsilon, \sigma_s^2, \gamma_{\theta_1}\}$ .

First, the noise variance can be estimated from the observed data by the following differencing technique (Von Neumann, 1941)

$$\widehat{\sigma_\epsilon^2} = \frac{1}{2n(p-1)} \sum_{i=1}^n \sum_{j=1}^{p-1} (Y_{i(j+1)} - Y_{ij})^2. \quad (6)$$

Hence an empirical estimate of the noise precision is given by  $\widehat{\gamma_\epsilon} = 1/\widehat{\sigma_\epsilon^2}$ .

Second, since we have the following formula for the covariance of the observed functional data (signal plus noise)

$$\text{Cov}(\mathbf{Y}_i) = \boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I} = \sigma_s^2 \mathbf{V} + \sigma_\epsilon^2 \mathbf{I},$$

from the *Signal + Noise* observation model (1), a moment estimator of  $\sigma_s^2$  can be derived as

$$\widehat{\sigma_s^2} = \frac{\text{trace}(\text{Cov}(\mathbf{Y}_i)) - p\widehat{\sigma_\epsilon^2}}{\text{trace}(\mathbf{V})} \approx \frac{\text{trace}(E[\text{Cov}(\mathbf{Y}_i)]) - p\widehat{\sigma_\epsilon^2}}{\text{trace}(E[\mathbf{V}])},$$

where  $\mathbf{V} \sim IW(\mathbf{A}, \delta)$ ,  $E[\mathbf{V}] = \mathbf{A}/(\delta - 2)$  and  $\text{trace}(\mathbf{A}) = p$ . Then the moment estimate of  $\sigma_s^2$  can be calculated by

$$\widehat{\sigma_s^2} = \frac{\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p (Y_{ij} - \bar{Y}_{\cdot j})^2 - p\widehat{\sigma_\epsilon^2}}{p/(m-p-1)}, \quad \bar{Y}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n Y_{ij}. \quad (7)$$

Third, an empirical estimate  $\widehat{\theta}_1$  can be obtained by matching the average of a selected off-diagonal of the sample correlation matrix (usually the one

with average closest to 0.5) with the value given by  $Matern(d; \theta_1, \theta_2 = 2.5, \sigma^2 = 1)$ , where  $d$  is determined by the chosen off-diagonal.

Finally, we set these empirical estimates as the expectations and choose a multiple of these estimates as the variances of the corresponding Gamma distributions of  $\{\gamma_\epsilon, \sigma_s^2, \gamma_{\theta_1}\}$ , and then solve for those hierarchical parameter values. For example, let  $Var(\gamma_\epsilon) = 3\hat{\gamma}_\epsilon$ , then  $(b_\epsilon = 1/3, a_\epsilon = b_\epsilon \hat{\gamma}_\epsilon)$  can be solved from equations  $\{Var(\gamma_\epsilon) = a_\epsilon/b_\epsilon^2 = 3\hat{\gamma}_\epsilon; E[\gamma_\epsilon] = a_\epsilon/b_\epsilon = \hat{\gamma}_\epsilon\}$ . Similarly, we will obtain values for  $(a_s, b_s, a_{\theta_1}, b_{\theta_1})$  by letting  $Var(\sigma_s^2) = \hat{\sigma}_s^2/30$  and  $Var(\gamma_{\theta_1}) = 3\hat{\gamma}_{\theta_1}$ .

The results of the case studies in Sections 4 and 5 suggested that this empirical Bayes approach performed very well.

## Acknowledgement

This research is supported by the NIH grant PO1-CA-082710. The authors want to thank all colleagues working on the PO1 project for collecting and processing the spectroscopy data.

## References

- Armstrong, H., Carter, C. K., Wong, K. F. K., and Kohn, R. (2009). “Bayesian covariance matrix estimation using a mixture of decomposable graphical models.” *Statistics and Computing*, 19(3): 303–316.
- Buys, T. P., Cantor, S. B., Guillaud, M., Adler-Storthz, K., Cox, D. D., Okolo, C., Arulogon, O., Oladepo, O., Basen-Engquist, K., Shinn, E., et al. (2012). “Optical technologies and molecular imaging for cervical neoplasia: a program project update.” *Gender medicine*, 9(1): S7–S24.
- Casella, G. (1985). “An introduction to empirical Bayes data analysis.” *The American Statistician*, 39(2): 83–87.
- Chan, J. C.-C. and Jeliazkov, I. (2009). “MCMC estimation of restricted covariance matrices.” *Journal of Computational and Graphical Statistics*, 18(2): 457–480.
- Chen, C.-F. (1979). “Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 235–248.

- Cuevas, A., Febrero, M., and Fraiman, R. (2002). “Linear functional regression: the case of fixed design and functional response.” *Canadian Journal of Statistics*, 30(2): 285–300.
- Daniels, M. J. and Kass, R. E. (1999). “Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models.” *Journal of the American Statistical Association*, 94(448): 1254–1263.
- Dawid, A. P. (1981). “Some matrix-variate distribution theory: notational considerations and a Bayesian application.” *Biometrika*, 68(1): 265–274.
- Falkner, F. and Falkner, F. T. (1960). *Child development: an international method of study*. Karger.
- Gardes, L., Girard, S., and Lekina, A. (2010). “Functional nonparametric estimation of conditional extreme quantiles.” *Journal of Multivariate Analysis*, 101(2): 419–433.
- Geman, S. and Geman, D. (1984). “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6: 721–741.
- Hall, P., Fisher, N. I., and Hoffmann, B. (1994). “On the nonparametric estimation of covariance functions.” *The Annals of Statistics*, 2115–2134.
- Hastings, W. K. (1970). “Monte Carlo sampling methods using Markov chains and their applications.” *Biometrika*, 57(1): 97–109.
- Hitchcock, D. B., Casella, G., and Booth, J. G. (2006). “Improved estimation of dissimilarities by presmoothing functional data.” *Journal of the American Statistical Association*, 101(473): 211–222.
- James, G. M., Hastie, T. J., and Sugar, C. A. (2000). “Principal component models for sparse functional data.” *Biometrika*, 87(3): 587–602.
- Jiang, C.-R. and Wang, J.-L. (2010). “Covariate adjusted functional principal components analysis for longitudinal data.” *The Annals of Statistics*, 38(2): 1194–1226.
- Matérn, B. et al. (1960). “Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations.” *Meddelanden fran statens Skogsforskningsinstitut*, 49(5).
- Müller, H.-G. (2005). “Functional Modelling and Classification of Longitudinal Data.” *Scandinavian Journal of Statistics*, 32(2): 223–240.

- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.  
URL <http://www.R-project.org/>
- Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.
- Ramsay, J. O. and Dalzell, C. (1991). “Some tools for functional data analysis.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 539–572.
- Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. Springer New York.
- Sampson, P. D. and Guttorp, P. (1992). “Nonparametric estimation of nonstationary spatial covariance structure.” *Journal of the American Statistical Association*, 87(417): 108–119.
- Silverman, B. W. (1996). “Smoothed functional principal components analysis by choice of norm.” *The Annals of Statistics*, 24(1): 1–24.
- Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer.
- Vieu, P. and Ferraty, F. (2006). *Nonparametric functional data analysis*. Springer Science+ Business Media.
- Von Neumann, J. (1941). “Distribution of the ratio of the mean square successive difference to the variance.” *The Annals of Mathematical Statistics*, 12(4): 367–395.
- Wand, M. M. P. and Jones, M. C. (1995). *Kernel smoothing*, volume 60. Crc Press.
- Whittle, P. (1954). “On stationary processes in the plane.” *Biometrika*, 434–449.
- Yao, F., Müller, H.-G., and Wang, J.-L. (2005). “Functional linear regression analysis for longitudinal data.” *The Annals of Statistics*, 33(6): 2873–2903.
- Zhu, H., Brown, P. J., and Morris, J. S. (2011a). “Robust, adaptive functional regression in functional mixed model framework.” *Journal of the American Statistical Association*, 106(495).
- Zhu, H., Dunson, D., and Strawn, N. (2011b). “Bayesian graphical models for multivariate functional data.” *Technical Report*.